FACCE MACSUR Mid-term Scientific Conference, »Achievements, Activities, Advancement«
Sassari, April 01-04, 2014

Interrelationship between evaluation metrics to assess agro-ecological models

Mattia Sanna[1*], Marco Acutis[1], Gianni Bellocchi[2]

[1]Department of Agricultural and Environmental Sciences - Production, Landscape, Agroenergy, University of Milan, Via Celoria 2, 20133 Milan, Italy

[2]Grassland Ecosystem Research Unit, French National Institute of Agricultural Research, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France

[*]Corresponding Author, e-mail address: mattia.sanna@unimi.it

Abstract— When evaluating the performances of simulation models, the perception of the quality of the outputs may depend on the statistics used to compare simulated and observed data. In order to have a comprehensive understanding of model performance, the use of a variety of metrics is generally advocated. However, since they may be correlated, the use of two or more metrics may convey the same information, leading to redundancy. This preliminary study intends to investigate the interrelationship between evaluation metrics, with the aim of identifying the most useful set of indicators, for assessing simulation performance. Our focus is on agro-ecological modelling. Twenty-one performance indicators were selected to compare simulated and observed data of three agronomic and meteorological variables: above-ground biomass, hourly air relative humidity and daily solar radiation. Indicators were calculated on large data sets, collected to effectively apply correlation analysis techniques. For each variable, the interrelationship between each pair of indicators was evaluated, by computing the Spearman's rank correlation coefficient. A definition of "stable correlation" was proposed, based on the test of heterogeneity, allowing to assess whether two or more correlation coefficients are equal. An optimal subset of indicators was identified, striking a balance between number of indicators, amount of provided information and information redundancy. They are: Index of Agreement, Squared Bias, Root Mean Squared Relative Error, Pattern Index, Persistence Model Efficiency and Modified Modelling Efficiency. The present study was carried out in the context of CropM-LiveM cross-cutting activities of MACSUR knowledge hub.

Index Terms— Model Evaluation, Performance indicators, Stable Correlation.

---

1 Introduction

Model evaluation is an essential step in the simulation process and one of the issues which has mostly interested the modelling community in the last years. The aggregation of multiple indicators of model performance into a single score offers a valuable way to assess models (after Bellocchi et al. 2002). So it is, because models performing well with respect to an indicator may not appear effective when evaluated using another indicator of performance (e.g. Rivington et al. 2005). We developed a systematic approach, for selecting the most suitable evaluation measures to assess the performances of dynamic simulation models. Our attention was focused on agro-ecological modelling, taking into account a broad set of commonly used performance indicators. The starting point of our study was the

analysis of the interrelationship between different evaluation metrics. Then, a correlation analysis was performed in which correlation coefficients were calculated between all resulting indicators from a test database of indicators. Subsequently, the definition of "stable correlation" was introduced, in order to identify any pattern in the correlations between the indicators. In the end, an optimal set of indicators was proposed, selecting those indicators, which, at the same time, are able to: 1) condense the greatest amount of information, showing a strong relationship with many of the other indicators; 2) be combined such that the condensed information provided by one of them do not overlap, but mutually complement each other. Preliminary results of the application of such approach are presented.

2 Materials and Methods

2.1 Data collection

A literature review was conducted to compile and classify an exhaustive list of indicators used to evaluate the performance of in agro-ecological models (Table 1).

Table 1 Classification of model performance indicators

Bias	Mean Bias Error (MBE), Squared Bias (SB), Fractional Bias (FB), Coefficient of Residual Mass (CRM)
Accuracy	Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Root Mean Squared Relative Error (RMSRE), Root Mean Squared Variation (RMSV), General Standard Deviation (GSD), Normalized Mean Squared Error (NMSE)
Efficiency	Modelling Efficiency (EF), Modified Modelling Efficiency (EF1)
Persistence	Persistence Model Efficiency (PME)
Correlation / Regression	Pearson's Correlation Coefficient (r), Spearman's Correlation Coefficient (r_s), Coefficient of Determination (r^2), Index of Agreement (d)
Median based	Robust Modelling Efficiency (REF), Median Absolute Error (MdAE), Relative Median Absolute Error (RMdAE)
Pattern	Pattern Index (PI)

A consistent data set of measured and estimated values was collected based on publicly available sources (literature and databases of regional meteorological services). Three of the most important agronomic and meteorological variables were used for this study, namely above-ground crop biomass over the growing season (AGB [t ha^{-1}]), hourly air relative humidity (HARH [%]) and daily solar radiation (RAD [MJ m^{-2}]), for which estimated values were available from simulation models. Series of estimated and measured data of above-ground rice biomass was extracted from Confalonieri & Bocchi (2005a, 2006). In this case, estimated data consist of simulation runs performed on 10 data sets with three crop

models: CropSyst (Stöckle et al. 2003), WARM (Confalonieri et al. 2005b)and WOFOST (Van Keulen & Wolf 1986). Estimated and measured values of hourly air relative humidity were selected among a large data set, which is part of an integrated evaluation with 13 modelling solutions (Bregaglio et al. 2010). Data from nine data sets were provided by the authors. Measured values of daily global solar radiation for 10 sites were collected from RAM Piemonte (Agrometeorological network of Piedmont Region) and ARPA Lombardia (Environmental Protection Agency of Lombardy Region) databases. Three simulation models were applied to estimate global solar radiation, namely Hargreaves (Hargreaves & Samani 1982), Bristow-Campbell (Bristow & Campbell 1984)and Campbell-Donatelli (Campbell & Donatelli, 1998).

2.2 Statistical analysis

The interrelationship between performance indicators was investigated by first computing the Spearman's rank correlation coefficient for each each pair of indicators and for each selected variable, obtaining three matrices (one for each variable). All the pairs were subsequently analysed to assess whether some stability could be detected in the correlation pattern, that is, if similar correlations can be found for all the three variables. The first step was to group the Spearman's correlation coefficients, corresponding to a given pair of indicators, in one set containing three values (Spearman's r values calculated on AGB, HARH and RAD data). The method for k independent samples proposed by Weaver & Wuensch (2013) was then applied to verify that population correlations were the same at p=0.05 level of significance. The entire procedure was repeated for each pair of performance indicators. As the "degree of stability" of any single correlation can be naturally expressed in terms of p values, a correlation value was defined as "stable" if its associated p value is greater than 0.05. This definition was introduced in order to meet the need of a criterion to assess possible relationships between performance indicators: if the correlation between two indicators is stable, it means that it does not depend on the type of data used; if the correlation does not depend on the type of data, it stands to reason that it is an intrinsic feature of the indicators themselves. Moreover, if a stable correlation between a pair of indicators is detected, then the correlation coefficients from a given combination of variables can be considered as equivalent and, consequently, their mean can be calculated without significant loss of information. This mean can be interpreted as a measure of the strength of the relationship between two performance indicators and, consequently, it can be used to quantify their shared information: the stronger the relationship, the greater the amount of common information provided by the two indicators. In the end, it was possible to select the most useful set of indicators, as follows: first, those showing a strong relationship with many other indicators are included, because they

somehow can condense a lot of information; secondly, they can be combined without overlap of the condensed information (rather, they mutually complement each other).

3 Results

Starting from the matrix of p-values for the heterogeneity tests, the mean values of the correlation coefficients were calculated for each pair of indicators exhibiting a stable correlation. In so doing, it was possible to propose an optimal subset of performance indicators, striking a balance between number of indicators, amount of information provided by each of them and information redundancy (Table 2).

Table 2.Mean values of correlation coefficients calculated on AGB, HARH and RAD data

Mean	MBE	FB	CRM	MSE	RMSE	RMSV	GSD	NMSE	EF	r	r_s	r^2	REF	MdAE	RmdAE
SB				0.79	0.79	0.49	0.69	0.73							
RMSRE									-0.53			-0.03	-0.60		
EF1															
PME							-0.85	-0.84	0.92						
d	-0.68	-0.69	0.69			0.20				0.15	0.18				
PI						0.70	0.56	0.61						0.53	0.55

First, d and SB were selected, because they are strongly related to indicators belonging to Bias, Correlation/Regression and Accuracy groups. Secondly, RSMRE and PI were chosen, being able to condense information from Efficiency and Median based groups. PME was assigned to the optimal subset, because no stable correlation was detected with any of the previous metrics. Moreover, PME shares a great amount of information with GSD, NMSE and EF, as shown by the mean values of the correlation coefficients. As no stable correlation was obtained between EF1 and any other metric, EF1 was also selected. It is important to observe that each performance indicator is strongly related to at least one element of the optimal subset. Moreover, in each column it is always possible to identify a mean value of correlation coefficients greater than 0.5 in absolute value, with the only exceptions of r, r_s and r^2.

4 Conclusions

The aim of this preliminary study was to shed some light on the interrelationship between different model evaluation metrics. A statistically-based approach was developed, introducing the concept of stable correlation to identify statistically equivalent correlation coefficients. Afterwards, the approach was applied on both meteorological and crop data, giving some insights about the interrelationships we were looking for. The performance indicators of the optimal subset are spread over the different groups

of Table 1, with the exception of the Median based group. Moreover, it is worth to underline that the presence of PME is strictly related to the time-varying nature of the selected dataset, since, in the PME case, the model prediction is compared relative to the performance of a persistence model where the model prediction at a given time step equals the observation at the previous time step. The proposed approach is also effective in minimizing information redundancy, since the indicators of the optimal subset overlap in very few cases. Such overlappings are clearly displayed in Table 2, as they occur where there is more than one value along a single column, meaning the corresponding performance indicator is correlated to more than one element of the optimal subset. These results are encouraging to identify some performance indicators that can be used together to increase confidence in model results with no redundancy, but they need to be confirmed for their stability on other variables and datasets. The proposed approach provides an objective criterion, as a consequence, it could be adopted other than in agro-ecological modelling. Moreover, once this procedure is applied on an adequate number of data sets and the existence of an optimal subset of indicators can be demonstrated, it can be considered as a sort of standard for model evaluation.

5 References

Alexandrov, G. A. et al., 2011. Technical assessment and evaluation of environmental models and software: Letter to the Editor. Environmental Modelling & Software, 26(3), pp. 328–336.

Bellocchi, G., Acutis, M., Fila, G. & Donatelli, M., 2002. An indicator of solar radiation model performance based on a fuzzy expert system. Agronomy Journal, 94, pp. 1222-1233.

Bellocchi, G., Rivington, M., Donatelli, M. & Matthews, K., 2010. Validation of biophysical models: issues and methodologies. Agronomy for Sustainable Development, 30, pp. 109-130.

Bregaglio, S. et al., 2010. An integrated evaluation of thirteen modelling solutions for the generation of hourly values of air relative humidity. Theoretical and Applied Climatology, 102, pp. 429–438.

Bristow, K. L. & Campbell, G. S., 1984. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agricultural and Forest Meteorology, 31(2), pp. 159–166.

Campbell, G. S. & Donatelli, M., 1998. A simple model to estimate global solar radiation. Proceedings of the 5th European Society of Agronomy Congress, Zima M., Bartosova M. eds., Nitra (Slovak), pp. 133-134.

Cochran, W. G., 1954. The combination of estimates from different experiments. Biometrics, 10, pp. 101-129.

Confalonieri, R. et al., 2005b. WARM: a scientific group on rice modelling. Italian Journal of Agrometeorology, Volume 2, pp. 54-60.

Confalonieri, R. & Bocchi, S., 2005a. Evaluation of CropSyst for simulating the yield of flooded rice in northern Italy. European Journal of Agronomy, 23, pp. 315-326.

Confalonieri, R., Gusberti, D., Bocchi, S. & Acutis, M., 2006. The CropSyst model to simulate the N balance of rice for alternative management. Agronomy for Sustainable Development, 26, pp. 241-249.

Diodato, N. & Bellocchi, G., 2007. Modelling solar radiation over complex terrains using monthly climatological data. Agricultural and Forest Meteorology, 144, pp. 111-126.

Hargreaves, G. H. & Samani, Z. A., 1982. Estimating potential evapotranspiration. Journal of Irrigation and Drainage Engineering, 108(IR3), pp. 223-230.

Huth, N. & Holzworth, D., 2005. Common sense in model testing. Proceedings of MODSIM 2005:

"International Congress on Modelling and Simulation". Modelling and Simulation Society of Australia and New Zealand Inc, December 2005, Melbourne (Australia), pp. 2804-2809.

Rivington, M., Bellocchi, G., Matthews, K. B. & Buchan, K., 2005. Evaluation of three model estimations of solar radiation at 24 UK stations. Agricultural and Forest Meteorology, 135, pp. 228-243.

Stöckle, C. O., Donatelli, M. & Nelson, R., 2003. CropSyst, a cropping systems simulation model. European Journal of Agronomy, 18, pp. 289-307.

Van Keulen, H. & Wolf, J., 1986. Modelling of agricultural production: weather, soils and crops. Wageningen: Pudoc.

Weaver, B. & Wuensch, K. L., 2013. SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. Behavior Research Methods, 45, pp. 880-895.